Dynamic Backdoor Attacks Against Machine Learning Models

Year: 2020 | Citations: 302 | Authors: A. Salem, Rui Wen, M. Backes, Shiqing Ma, Yang Zhang

Abstract

Machine learning (ML) has made tremendous progress during the past decade and is being adopted in various critical real-world applications. However, recent research has shown that ML models are vulnerable to multiple security and privacy attacks. In particular, backdoor attacks against ML models have recently raised a lot of awareness. A successful backdoor attack can cause severe consequences, such as allowing an adversary to bypass critical authentication systems. Current backdooring techniques rely on adding static triggers (with fixed patterns and locations) on ML model inputs which are prone to detection by the current backdoor detection mechanisms. In this paper, we propose the first class of dynamic backdooring techniques against deep neural networks (DNN), namely Random Backdoor, Backdoor Generating Network (BaN), and conditional Backdoor Generating Network (c-BaN). Triggers generated by our techniques can have random patterns and locations, which reduce the efficacy of the current backdoor detection mechanisms. In particular, BaN and c-BaN based on a novel generative network are the first two schemes that algorithmically generate triggers. Moreover, c-BaN is the first conditional backdooring technique that given a target label, it can generate a target-specific trigger. Both BaN and c-BaN are essentially a general framework which renders the adversary the flexibility for further customizing backdoor attacks. We extensively evaluate our techniques on three benchmark datasets: MNIST, CelebA, and CIFAR-10. Our techniques achieve almost perfect attack performance on back-doored data with a negligible utility loss. We further show that our techniques can bypass current state-of-the-art defense mechanisms against backdoor attacks, including ABS, Februus, MNTD, Neural Cleanse, and STRIP.